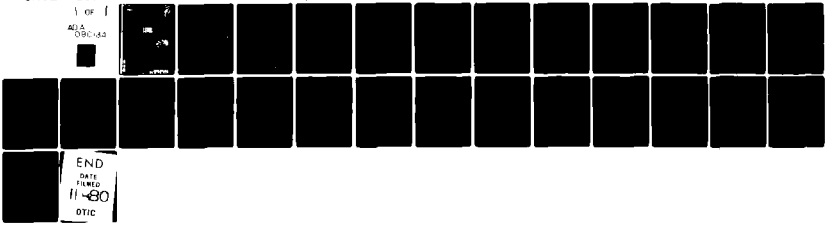
AD-A090 134 ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF MATHEMATICS F/G 12/1
CONTROLLED PROBABILITY PROPORTIONAL TO SIZE SAMPLING DESIGNS.(U)
JUL 80 A HEDAYAT, B Y LIN AFOSR-76-3050

UNCLASSIFIED NL



END
DATE FILMED
11-80
DTIC

(6)

CONTROLLED PROBABILITY PROPORTIONAL TO SIZE SAMPLING DESIGNS

(4)

by

(9) Interim rept.

A./Hedayat and B.Y./Lin
Department of Mathematics
University of Illinois at Chicago Circle

(12) 27

July 1980

16 19769

111-1

(15) Research is supported by Grant AFOSR-76-3050

CONTROLLED PROBABILITY PROPORTIONAL TO SIZE SAMPLING DESIGNS*

By A. Hedayat and B.Y. Lin

Department of Mathematics

University of Illinois, Chicago

ABSTRACT

Any sampling design, d, of size n without replacement based on a finite population U of N units or strata can be formally presented by a pair (S_d, P_d), where S_d called the support of d is any set of subsets of size n each based on the elements of U such that the (set theoretic) union of these subsets, called samples, is U and P_d is a strictly positive probability distribution on S_d. A sampling design is said to be a probability proportional to size, denoted by PPS(N,n), if the probability that the unit i is being selected in a random sample is proportional to a known positive quantity associated with the unit i = 1,2,...,N. The literature of survey sampling offers a PPS(N,n) with S_d consists of all $\binom{N}{n}$ possible samples. Here we give an easily applicable technique for the construction of PPS(N,n) with various support sizes and various probabilities on each support. Such sampling designs are needed for controlled samplings when some samples are undesirable to be chosen or we need to minimize (or maximize) the probabilities of the selection of certain samples.

*Research is supported by Grant AFOSR76-3050C

CONTROLLED PROBABILITY PROPORTIONAL TO SIZE SAMPLING DESIGNS

By A. Hedayat and B.Y. Lin

1. Introduction Let U be a finite population of N units or N strata.

A sampling design without replacement, d, based on U is a pair (S_d, P_d) where S_d, called the support of d, is any set of nonempty subsets of U and P_d is a strictly positive probability distribution on S_d with

$$\bigcup_{s_d \in S_d} s_d = U \tag{1.1}$$

Every member of S_d is called a sample and a random sample (or probability sample) is a sample selected by implementing d. In general samples in S_d may have different sizes.

In order to implement d we must know the precise structures of S_d and P_d. However, for the purpose of customary statistical analysis of the data collected via a random sample s_d all we need to know are the following quantities called respectively the first order and the second order inclusion probabilities:

$$\Pi_{di} = \text{prob. that a random sample will contain the unit } i = \sum_{s_d \ni i} p_d(s_d) \tag{1.2}$$

$$\Pi_{dij} = \text{prob. that a random sample will contain the units } i \text{ and } j \neq i = \sum_{s_d \ni i,j} p_d(s_d). \tag{1.3}$$

Note that by (1.1), $\Pi_{di} > 0$. However, Π_{dij} can be zero for some i and j. Indeed some of the classical sampling designs, such as systematic samplings, have the undesirable property that $\Pi_{dij} = 0$ for some i and j. If we are interested in unbiased estimation of variances of linear estimators we should avoid such sampling designs.

In the context of our discussion the literature of survey sampling are basically of two types:

1. Those which do not specify S_d and P_d but rather give procedures for drawing

random samples. Among these some specify Π_{di}'s and Π_{dij}'s and some give only Π_{di}'s and leave the burden of deriving Π_{dij}'s to the reader. Most papers are of this latter type.

2. Those which specify S_d and P_d. The values of Π_{di}'s and Π_{dij}'s are either given or can be easily computed by knowing S_d and P_d. Unfortunately, only few papers are of this type.

2. PPS Sampling A purpose of survey sampling is to study a characteristic of interest by a random sample chosen via a sampling design. Let us denote this characteristic of interest by y. There is often a case that besides the characteristic y there exists some auxiliary characteristic x which is related to y and its information is available to us. So, to each unit i in U there are associated two measurements X_i(known to us) and Y_i corresponding to the characteristics x and y respectively. We want to estimate the population total $Y = Y_1+Y_2+\ldots+Y_N$ utilizing the information provided by a random sample generated by a sampling design d. It is known (see the list of papers at the end) that in some cases we can improve the precision of our estimator of Y if we properly utilize the information provided by X_i's in the formation of the sampling design. One such sampling design, popular among survey statisticians, is called probability proportional to size (PPS) sampling design. Through our notation this is defined as:

Definition 2.1 A sampling design, $d = (S_d,P_d)$, based on U is called a probability proportional to size design of size n, designated by PPS(N,n), if (i) each sample in S_d consists of n units, and (ii) the probability Π_{di} is proportional to q_i (hence the name), where $q_i = X_i/\sum_{j=1}^{N} X_j$.

Since in any sampling design based on samples of size n

$$(2.1)\quad \sum_{i=1}^{N} \Pi_{di} = \sum_{i=1}^{N} \sum_{s_d \ni i} P_d(s_d) = n,$$

therefore in PPS(N,n) sampling design

$$(2.2)\quad \Pi_{di} = nq_i$$

which puts the requirement $nq_i<1$ for the existence of such designs. As we shall see later without further demands on d such sampling designs always exist. Here we would like to emphasize two points. First, q_i does not have to be of the form $X_i/\sum_{j=1}^{N} X_j$. In general, we allow q_i, referred to as the size of the unit i, to be any positive number as long as $q_1+q_2+\ldots+q_N = 1$ and $nq_i<1$. However, for all practical purposes we can assume q_i to be a rational number. Second, we should take advantage of the mild requirements of PPS(N,n) in preparing a sampling design which meets other useful requirements. We shall explain in details this latter point later on.

In the literature of survey sampling, PPS sampling designs belong to a celebrated family of sampling designs known as unequal probability sampling designs without replacement. Our purpose here is not to review the literature on this family but rather to indicate where and how our contributions fit in the literature dealing primarily with PPS sampling designs. The interested reader on the subject of unequal probability sampling designs should consult the selected bibliography and their corresponding references at the end of the paper.

The first formal attempt to construct PPS(N,n) sampling design was undertaken by Goodman and Kish(1950). These authors do not specify S_d or P_d but rather give a procedure for drawing a random sample which guarantees the proportionality of Π_i to q_i. It is extremely difficult to derive a general expression for Π_{ij}'s of the procedure of Goodman and Kish since the mathematical structure of their procedure is quite complicated. Hartley and Rao(1962) used asymptotic theory to approximate Π_{ij}'s of the procedure of Goodman and Kish. Though it is difficult to specify P_d of the corresponding design of Goodman and Kish it is easy to see that S_d consists of all possible $\binom{N}{n}$ samples each receiving positive probability. Brewer(1963) and Durbin(1967) were able to construct PPS(N,2) sampling designs. Their designs have the

property that $\Pi_{dij} > 0$ and therefore S_d consists of all possible $\binom{N}{2}$ samples. Sampford(1967) inspired by the results of Brewer(1963) and Durbin(1967) was able to construct PPS(N,n) sampling designs for all N and n. Again the support of Sampford's design is all $\binom{N}{n}$ possible samples. Sampford designs have the desirable properties that $\Pi_{ij} > 0$ and $\Pi_{ij} < \Pi_i \Pi_j$. The statistical usefulness of these latter properties can be argued as follows. To estimate the population total Y we can use the Horvitz-Thompson linear unbiased estimator

$$\hat{Y}_{HT} = \sum_{i \in s_d} y_i/\Pi_{di} \tag{2.3}$$

It can be easily verified that

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\Pi_{di}\Pi_{dj} - \Pi_{dij})\left(\frac{Y_i}{\Pi_{di}} - \frac{Y_j}{\Pi_{dj}}\right)^2 \tag{2.4}$$

and can be unbiasedly estimated by the Yates-Grundy estimator

$$\hat{Var}(\hat{Y}_{HT}) = \sum_{i \neq j \in s_d}^{n} \sum^{n} \frac{\Pi_{di}\Pi_{dj} - \Pi_{dij}}{\Pi_{dij}} \left(\frac{y_i}{\Pi_{di}} - \frac{y_j}{\Pi_{dj}}\right)^2 \tag{2.5}$$

provided $\Pi_{dij} > 0$. In practice it is desirable that this estimator be always nonnegative. The property that $\Pi_{dij} < \Pi_{di}\Pi_{dj}$ guarantees this.

The additional properties that $\Pi_{dij} > 0$ force the size of the support of PPS(N,n) to be $\binom{N}{n}$ if n = 2. However, for n>2 it is possible to construct PPS(N,n) sampling designs whose supports have less than $\binom{N}{n}$ samples in them. This allows us to put zero probability on samples which we consider to be undesirable or uneconomical to collect data from. The published literature provides no such opportunities. We are also able to construct for the given N,n and sizes $q_1, q_2, \ldots, q_N$ various S_d with varieties of P_d. Again here we are able to control the members of S_d and their corresponding probabilities. This means that either we can exclude undesirable samples from S_d or put very little probabilities on them for the selection purposes. If we are interested in all $\binom{N}{n}$ samples then our procedure allows to construct various P_d in contrast to the procedure of Sampford which provides no choice at all. Before we close this section we give two examples to elucidate the above points.

Example 2.1 Let U be a stratum of size N = 5 with the following sizes: $q_1 = 3/11$, $q_2 = 1/11$, $q_3 = 2/11$, $q_4 = 3/11$ and $q_5 = 2/11$. Suppose we want to select a sample of size n = 2 by the method of PPS sampling. Further, we desire that $\Pi_{ij} > 0$, and $\Pi_{ij} < \Pi_i \Pi_j$ so that the Yates-Grundy estimator(2.5) does not take negative value. Since n = 2 and we require that $\Pi_{ij} > 0$ any PPS sampling should have all $\binom{5}{2} = 10$ samples of size 2 in the support. So we have no problem concerning the construction of the support. Therefore, all we have to do is to specify probabilities on these samples. For the given N, n, $q_1, q_2, \ldots, q_N$ Sampford(1967) gives one such a set of probabilities. Since in this case n = 2 the Sampford probabilities are identical to those of Durbin(1967). Our procedure (see Section 3) gives several such possibilities. Below we list two such choices. Thus in this case we could not control the support due to the restrictions imposed on the design but we could control the probabilities on the samples.

S_d: Support	P_d: Probability on the Support		
(samples)	Durbin/Sampford design	Examples of our designs 1	2
12	147/2497	1/22	1/22
13	324/2497	3/22	3/22
14	567/2497	6/22	5/22
15	324/2497	2/22	3/22
23	80/2497	1/22	1/22
24	147/2497	1/22	1/22
25	80/2497	1/22	1/22
34	324/2497	2/22	3/22
35	180/2497	2/22	1/22
45	324/2497	3/22	3/22

Example 2.2 Suppose we have a stratum of N = 6 units and would like to select 3 units based on a PPS sampling with the following sizes: $q_1 = 2/17$, $q_2 = 3/17$, $q_3 = 4/17$, $q_4 = 1/17$, $q_5 = 2/17$, $q_6 = 5/17$. Let us also require that $\Pi_{ij} > 0$ and $\Pi_{ij} < \Pi_i\Pi_j$. As we said the only available procedure in the published literature is that of Sampford(1967). In the following table we give the Sampford design as well as a design generated by our procedure.

Sampford design		Our design	
S_d:support (samples)	P_d:probability on each sample	S_d:support (samples)	P_d:probability on each sample
123	44352/1529898	123	1/34
124	5445/1529898	126	3/34
125	12600/1529898	136	7/34
126	121275/1529898	145	1/34
134	10560/1529898	234	1/34
135	24192/1529898	235	1/34
136	221760/1529898	236	7/34
145	2880/1529898	246	2/34
146	29700/1529898	256	3/34
156	67200/1529898	346	1/34
234	19602/1529898	356	6/34
235	44352/1529898	456	1/34
236	381150/1529898		
245	5445/1529898		
246	54450/1529898		
256	121275/1529898		
345	10560/1529898		
346	101640/1529898		
356	221760/1529898		
456	29700/1529898		

Our design has all the desirable properties which the Sampford's design has. In addition, our design puts zero probability on the following eight samples 124, 125, 134, 135, 146, 156, 245, 345. Therefore, our sampling design could be utilized for controlled sampling if our desire is not to select these samples. So, in this case, we controlled the support of the sampling design.

The technique which we shall present in Section 3 can be adjusted to accommodate certain "reasonable" demands on the composition of S_d or structure of P_d. However, we do not like to leave the impression that:
(i) we can freely choose the samples in the support; or
(ii)we can arbitrarily manipulate the probabilities on the samples.

Clearly these cannot be done. For examples, the demand that $\Pi_{ij} > 0$ for all i and j, $i \neq j$ puts an obvious restriction on the composition of S_d and its cardinality, i.e., the samples in S_d must form a cover for all pairs which in turn puts a lower bound on the number of samples in S_d. Or, we cannot ask for a sampling design which puts zero probabilities on certain undesirable samples. What our technique is capable of doing is to minimize such probabilities though in some situations can indeed exclude such samples from the support.

3. Construction of Controlled PPS Sampling Designs. As we pointed out in Section 2 we can utilize Sampford's technique to construct a PPS(N,n) sampling design for every population size N, sample size n and any set of admissible unit sizes $q_1, q_2, \ldots, q_N$. While the technique of Sampford is an interesting one it is not applicable at all if we want to construct controlled PPS(N,n) sampling designs. In this section we shall provide a very general technique for the construction of such sampling designs. Our technique has no similarity to the technique of Sampford. Moreover, our technique enjoys the following practically useful features.(i) It is an easy technique to be understood and to be utilized in practical situations. (ii) It is a very flexible technique in a sense that we can adjust it to produce desirable sampling designs. For example, for given N, n and admissible unit sizes $q_1, q_2, \ldots, q_N$ it is possible

to adjust the technique to produce many PPS sampling designs with various support sizes and various probabilities on the samples in each support. This flexibility allows us to construct controlled PPS(N,n) sampling designs.

We shall now establish a result which is needed for the development of our technique to follow. Suppose we have N nonempty boxes containing $k_1, k_2, \ldots, k_N$ objects. For a given integer $n \leq N$ a round of size n is defined to be a process by which we select n boxes and remove one object from each box. Now the problem is this: what are the necessary and sufficient conditions on $N, n, k_1, \ldots k_N$ so that all the objects can be removed from the N boxes by a series of successive rounds of size n? This problem is completely solved in the following lemma.

Lemma 3.1. The necessary and sufficient conditions for removing $k_1+k_2+\ldots+k_N = M$ objects from N boxes by a series of successive rounds of size n are:

(1) $M \equiv 0 (\text{mod } n)$;

(2) $\max_i k_i \leq (M/n)$.

Proof. Necessity. (1) In each round of size n we remove n objects so the total number of objects, M, must be a multiple of n. (2) It takes precisely M/n rounds of size n to pick all the M objects. Therefore, no k_i can exceed the total number of rounds M/n.

Sufficiency. Consider the following procedure. At round one we remove one object from each of the n boxes containing the largest number of objects. Similarly, we proceed with the remaining (M/n)-1 rounds. We claim that this procedure will succeed in removing all the M objects as long as M is a multiple of n and $k_i \leq M/n$, $i=1,2,\ldots,N$. To simplify the proof we shall distinguish two distinct cases:

Case 1. There are precisely n boxes each containing M/n objects.

Case 2. There are less than n boxes each containing M/n objects.

Note that there cannot be more than n boxes each containing M/n objects.

Also, N=n in Case 1 and N >n in Case 2. It is clear that our procedure will succeed in Case 1. In Case 2 we shall establish that our procedure will end up to a case similar to Case 1 for the reduced values of M, and k_i's and thus we will be able to remove all the M objects. To show this we shall prove two things: First, we shall prove that at no time the reduced values of M and k_i's contradict the necessary conditions (1) and (2) of the lemma. After the completion of round j let $k_i^{(j)}$ and M_j be the number of objects left in the ith box and the total number of objects left in N boxes respectively.

Now we claim that:

$(1')$ $M_j \equiv 0 \pmod{n}$

$(2')$ $\max_i k_i^{(j)} \leq (M_j/n)$

$(1')$ is obvious by assumption (1) and the fact that $M_j = M-j(n)$.

(2') can be argued as follows. By assumption (2) $\max_i k_i \leq M/n$. Thus at the end of round j of our procedure $\max_i k_i^{(j)} \leq \max_i k_i - j$ which yields $\max_i k_i^{(j)} \leq (M/n) - j = M_j/n$. As we can see conditions (1') and (2') are equivalent to conditions (1) and (2) of the lemma for integers n, M_j and $k_1^{(j)}, k_2^{(j)},\ldots,k_N^{(j)}$. Therefore, at the beginning of each round the system satisfy the necessary conditions. Second, at the (j+1)th round we will be faced with two possibilities. N-n boxes will be empty and each of the remaining n boxes contains the same number of objects. As we pointed out in Case 1 above our procedure will clearly succeed. Otherwise, we will continue the rounds. If a situation as above never arises before round (M/n)-1 then round (M/n)-1 will produce a situation as above with one object in n boxes and thus by round (M/n) all the objects will be removed. Note that the conditions of the lemma will exclude the possibliity of ending up with more than N-n boxes to be empty at any round.

Remark 3.1. If in addition to conditions (1) and (2) above N is a multiple of n then it is possible that we end up in some round with all N boxes containing the same number of objects. Clearly, we can go on with our procedure in such a situation in a trivial manner.

The following example will elucidate the procedure outlined in the lemma and the point mentioned in the above remark.

Example 3.1. Consider the following system: N=6, n=3, k_1=6, k_2=9, k_3=12, k_4=3, k_5=6 and k_6=15. Here M=51 and M/n = 17. The necessary conditions are satisfied. The following 17 rounds of size 3 will remove all the 51 objects.

Box No.	1	2	3	4	5	6
no. of objects: k_i	6	9	12	3	6	15
Round 1		1	1			1
Residuals	6	8	11	3	6	14
Round 2		1	1			1
Residuals	6	7	10	3	6	13
Round 3		1	1			1
Residuals	6	6	9	3	6	12
Round 4	1		1			1
Residuals	5	6	8	3	6	11
Round 5			1		1	1
Residuals	5	6	7	3	5	10
Round 6		1	1			1
Residuals	5	5	6	3	5	9
Round 7	1		1			1
Residuals	4	5	5	3	5	8
Round 8		1			1	1
Residuals	4	4	5	3	4	7
Round 9	1		1			1

Residuals	3	4	4	3	4	6
Round 10		1	1			1
Residuals	3	3	3	3	4	5
Round 11	1				1	1
Residuals	2	3	3	3	3	4
Round 12		1		1		1
Residuals	2	2	3	2	3	3
Round 13			1		1	1
Residuals	2	2	2	2	2	2
Round 14	1			1	1	
Residuals	1	2	2	1	1	2
Round 15		1	1			1
Residuals	1	1	1	1	1	1
Round 16			1	1		1
Residuals	1	1	0	0	1	0
Round 17	1	1			1	
Residuals	0	0	0	0	0	0

Remarks 3.2. We would like to make the following important observations in the context of Example 3.1. (i) Except in rounds 1,2,3, 6,13, and 15 we had more than one choice in selecting $n = 3$ boxes. (ii) we could modify our procedure and empty boxes 3,5, and 6 in rounds 13, 14 and 15 and boxes 1,2, and 4 in rounds 16 and 17.(iii) After rounds 13 and 15 we ended up with two examples of the case pointed out in Remark 3.1 Again note that in a situation like this we could easily modify our procedure and empty the boxes in several ways.

The above example clearly demonstrates that in general there are many options in the formation of rounds and our procedure could be easily modified throughout the process. These properties are important when we apply our procedure in sampling from finite populations.

We shall now apply Lemma 3.1 and prove the following theorems. In Theorem 3.1 we shall give a technique for the construction of PPS(N,n) sampling designs directly based on the procedure of Lemma 3.1. In Theorem 3.2 we shall show how we can explicitly construct PPS(N,n) sampling designs with the added property that $\Pi_{ij} \neq 0$. Examples are given to demonstrate the techniques.

We recall from Section 2 that the unit sizes $q_1,q_2,\ldots,q_N$ in a PPS sampling should satisfy $q_i > 0$, $nq_i < 1$ and $q_1+q_2+\ldots+q_N = 1$. Here for all practical purposes we shall assume that all q_i's are in rational forms

Theorem 3.1. For any N, $n < N$, and unit sizes $q_1,q_2,\ldots,q_N$ there exists at least one PPS(N,n) sampling design.

Proof (By construction). Associate with the ith unit the integer $k_i = nq_iq$. Now pretend that the N units are N boxes with the ith box containing k_i objects. The N integers $k_1,k_2,\ldots,k_N$ and the sample size n clearly satisfy condition (1) of lemma 3.1. They also satisfy condition (2) since by assumption $nq_i < 1$ and thus $k_i < q = M/n$ for $M = k_1+k_2+\ldots+k_N$. Now by M/n rounds of size n empty these N boxes and keep a record of all rounds as we did in Example 3.1. Now our PPS(N,n) sampling design is defined as follows:

S_d: The Support. The set of n units in each round determines a sample in S_d. The (set theoretic) union of all these samples constitutes the support. Note that since there are $(M/n) = q$ rounds in all, thus the cardinality of $S_d \leq q$.

P_d: The Probability On The Support. If $S_d = \{i_1,i_2,\ldots,i_n\}$ is a sample in S_d then the probability on this sample, $p_d(s_d)$, is the proportion of rounds which produced this sample. Thus

$$p_d(s_d) = r(s_d)/q$$

where $r(s_d)$ is number of rounds of size n in which the units $i_1,i_2,\ldots,i_n$ were chosen.

Indeed, $d = (S_d, P_d)$ so defined is a PPS(N,n) sampling design. Three things should be verified. (a) The union of samples in S_d should satisfy condition (1.1). (b)We should show that P_d is a strictly positive probability distribution on S_d. (c) $\Pi_{di} = nq_i$, $i = 1, 2, \ldots, N$.

(a) is obvious since $q_i > 0$ and thus $k_i > 0$ meaning that there is at least one sample (one round) which contains the ith unit.

(b) clearly $p_d(s_d) = r(s_d)/q$ is a positive number less than one and

$$\sum_{s_d \in S_d} p_d(s_d) = \frac{1}{q} \sum_{s_d \in S_d} r(s_d) = \frac{1}{q} \cdot (\text{no. of rounds} = M/n) = 1$$

(c) for all i,

$$\Pi_{di} = \sum_{s_d \ni i} p_d(s_d) = \frac{1}{q} \sum_{s_d \ni i} r(s_d) = \frac{1}{q} \; k_i = nq_i.$$

Example 3.2. Let $N = 6$, $n = 3$ and $q_1 = 2/17$, $q_2 = 3/17$, $q_3 = 4/17$, $q_4 = 1/17$, $q_5 = 2/17$, $q_6 = 5/17$. In this case $q = 17$ and thus $k_1 = 6$, $k_2 = 9$, $k_3 = 12$, $k_4 = 3$, $k_5 = 6$ and $k_6 = 15$. We have already exhibited a table of rounds in Example 3.1 for this problem. So, let us exhibit the corresponding PPS(6,3) sampling design. For example, the 6 rounds, 1,2,3,4,9, 13 determine the sample $s_1 = \{2,3,6\}$ with $p_d(s_1) = 6/17$ and similarly the rest of the samples and the corresponding probabilities.

sample	probability	rounds produced the sample
2 3 6	6/17	1, 2, 3, 6, 10, 15
1 3 6	3/17	4, 7, 9
3 5 6	2/17	5, 13
2 5 6	1/17	8
1 5 6	1/17	11
2 4 6	1/17	12
1 4 5	1/17	14
3 4 6	1/17	16
1 2 5	1/17	17

Let us for example compute Π_2. There are 4 samples in the support which contain unit 2. If we add up the probabilities over these 4 samples we obtain $\Pi_2 = 6/17+1/17+1/17+1/17 = 9/17$ which is equal to $nq_2 = 3(3/17)$. Also note that in this design $\Pi_{ij} > 0$ for all $i, j \neq i$. Finally, our design has excluded 11 samples out of $\binom{6}{3}$ for sampling purposes.

Let us now look at the procedure outlined in Theorem 3.1 from other viewpoints. The only demand we formally imposed on the procedure in Theorem 3.1 was Π_i to be proportional to q_i as was specified in the definition of PPS sampling designs. Otherwise, we left the procedure very flexible so that we can adjust or modify it to produce desirable PPS samplings. If, for example, we further demand that $\Pi_{ij} > 0$ for all $i,j \neq i$ then we should adjust the procedure by observing the following facts.

Proposition 3.1. It is necessary that

$$\min_i k_i > \left\{\frac{N-1}{n-1}\right\}, \tag{3.1}$$

in order that a resulting sampling design generated by the procedure of Theorem 3.1 have further properties that $\Pi_{ij} > 0$ for all $i, j \neq i$. $\{z\}$ denotes the smallest integer greater than or equal to z.

Proof. The unit i appears in precisely k_i rounds. Thus the number of samples in the support which contain the unit i is at most k_i. Since unit i appears with $n-1$ other units in each sample, thus in order that $\Pi_{ij} > 0$ for all $j \neq i$ the ith unit should appear in at least $\{(N-1)/(n-1)\}$ samples in the support. Therefore, it is necessary that $k_i > \{(N-1)/(n-1)\}$ for $i = 1,2,\ldots,N$.

Now we shall show that in case k_i's do not satisfy condition (3.1) we can artificially increase the values of k_i's so that there will be enough samples in the support to cover all pairs of units. In some cases it may be necessary to manipulate the values of k_i's even though the k_i's satisfy the necessary condition (3.1). However, we recommend that this

device should be avoided if possible if we are interested in supports with not too many samples. Let us, for example, reconsider Example 3.1. In that example had we chosen boxes 1,2,3 in round 16 and consequently boxes 4,5,6 in round 17 the resulting sampling design would have suffered from the undesirable property that $\Pi_{34} = 0$.

To avoid such outcomes we should keep track of the pairs being covered as we go along and forming the rounds. We should take advantage of those situations in which we have several possibilities for the formation of rounds. In such a situation we should select a round which help in covering uncovered pairs by the preceeding rounds as we did in Example 3.1. As we mentioned above in any case we can artificially increase the values of k_i's to make sure that enough samples are in the support to cover all the pairs. We shall now explicitly indicate how to increase the values of k_i's without putting too many samples in the support. The procedure is applicable whether or not the k_i's satisfy the necessary condition(3.1). However, we shall explain it in the context of the case in which condition (3.1) is violated. Let $\min_i k_i = k^*$ and assume that $k^* < \{(N-1)/(n-1)\}$. Proceed as in procedure of Theorem 3.1 till the stage in which, $\bar{k}_i$'s, the reduced values of k_i's are very close to k^*.(We do not need to be too formal and introduce a measure of closeness since,as we shall see,the operation we shall apply can be introduced at any round, even in round one.) Multiply all $\bar{k}_i$'s by a sufficiently large integer h so that

$$h(\min_i \bar{k}_i) > \{(N-1)/(n-1)\}$$

and go on with the remaining rounds with these artificially large remaining $\hat{k}_i = h\,\bar{k}_i$, $i = 1,2,\ldots,N$. It is clear that if we select h large enough we can cover all the pairs (i,j). The reason we do not recommend this operation in round one,or early after that, is to avoid the prolongation of the procedure and consequently having too many unwanted samples in the support. There is a slight modification in forming the corresponding PPS sampling. The

support, as in *Theorem* 3.1, *consists of those samples formed* by the rounds. And for probabilities if s_d is a sample then the probability over it is computed by

$$p_d(s_d) = r^*(s_d)/q^*, \quad q^* = \sum_{s_d} r^*(s_d)$$

where, $r^*(s_d) = h(r_1(s_d)) + r_2(s_d)$ with

$r_1(s_d)$ = no. of rounds which produced s_d before the application of h;

and

$r_2(s_d)$ = no. of rounds which produced s_d after the application of h.

Now we give an example to explain the above ideas.

Example 3.3. Let N = 8, n = 3 and the unit sizes q_i's as given below:

unit	1	2	3	4	5	6	7	8
q_i	2/18	3/18	1/18	5/18	1/18	2/18	1/18	3/18
$3q_i$	6/18	9/18	3/18	15/18	3/18	6/18	3/18	9/18
k_i	6	9	3	15	3	6	3	9

Here $\min_i k_i = 3 < \{\frac{8-1}{3-1}\} = 4.$

unit	1	2	3	4	5	6	7	8
k_i	6	9	3	15	3	6	3	9
Rounds 1,2,3		3		3				3
Residuals	6	6	3	12	3	6	3	6
Round 4	1	1		1				
Residuals	5	5	3	11	3	6	3	6
Round 5				1		1		1
Residuals	5	5	3	10	3	5	3	5
Round 6		1		1		1		
Residuals	5	4	3	9	3	4	3	5
Round 7	1			1				1
Residuals	4	4	3	8	3	4	3	4
Round 8		1		1				1
Residuals	4	3	3	7	3	4	3	3
Round 9	1			1		1		
Residuals	3	3	3	6	3	3	3	3
Introduce h = 2	6	6	6	12	6	6	6	6

Round 10			1	1	1			
Residuals	6	6	5	11	5	6	6	6
Round 11				1			1	1
Residuals	6	6	5	10	5	6	5	5
Round 12	1	1		1				
Residuals	5	5	5	9	5	6	5	5
Round 13				1		1	1	
Residuals	5	5	5	8	5	5	4	5
Round 14	1		1	1				
Residuals	4	5	4	7	5	5	4	5
Round 15				1	1	1		
Residuals	4	5	4	6	4	4	4	5
Round 16		1		1				1
Residuals	4	4	4	5	4	4	4	4
Round 17	1			1	1			
Residuals	3	4	4	4	3	4	4	4
Round 18		1		1			1	
Residuals	3	3	4	3	3	4	3	4
Round 19			1			1		1
Residuals	3	3	3	3	3	3	3	3
Round 20	1		1				1	
Residuals	2	3	2	3	3	3	2	3
Round 21				1		1		1
Residuals	2	3	2	2	3	2	2	2
Round 22		1	1		1			
Residuals	2	2	1	2	2	2	2	2
Round 23					1		1	1
Residuals	2	2	1	2	1	2	1	1
Round 24	1	1		1				

Residuals	1	1	1	1	1	2	1	1
Round 25	1	1				1		
Residuals	0	0	1	1	1	1	1	1
Round 26			1	1	1			
Residuals	0	0	0	0	0	1	1	1
Round 27						1	1	1
Residuals	0	0	0	0	0	0	0	0

Note that we increased the values of the residuals $\bar{k}_i$'s at the end of round 9 in which these values were close to $\min_i k_i = 3$ to begin with. Since in this case h = 2 each round before round 10 is counted twice in computing the probabilities. The resulting PPS(8,3) with all $\Pi_{ij} > 0$ is given below.

sample	probability	sample	probability	sample	probability
2 4 8	9/36	4 7 8	1/36	3 6 8	1/36
1 2 4	4/36	4 6 7	1/36	1 3 7	1/36
4 6 8	3/36	1 3 4	1/36	2 3 5	1/36
2 4 6	2/36	4 5 6	1/36	5 7 8	1/36
1 4 8	2/36	1 4 5	1/36	1 2 6	1/36
1 4 6	2/36	2 4 7	1/36	6 7 8	1/36
3 4 5	2/36				

This PPS sampling has excluded $\binom{8}{3} - 19 = 37$ samples from the support.

We now summarize in Theorem 3.2 what we have discovered above.

Theorem 3.2. For any N,n < N and unit sizes $q_1, q_2, \ldots, q_N$ there are various probability proportional to sampling designs with various support sizes and varieties of probabilities on each support with $\Pi_{ij} > 0$, for all i, j ≠ i. These sampling designs could be used for the purpose of controlled sampling.

In conclusion, we would like to point out that as long as k_i's, or $k_i^{(j)}$'s, satisfy the conditions of Lemma 3.1 our procedure will succeed. So we do not have to choose the largest n boxes at every round. This fact together with the technique of introducing a multiplier makes our procedure

even more flexible. The following example should demonstrate the point.

Example 3.4. As in Example 3.2, let N = 6, n = 3 and the unit sizes q_i's as follows:

unit	1	2	3	4	5	6
q_i	2/17	3/17	4/17	1/17	2/17	5/17
$3q_i$	6/17	9/17	12/17	3/17	6/17	15/17
k_i	6	9	12	3	6	15

unit	1	2	3	4	5	6
k_i	6	9	12	3	6	15
Introduce h = 2	12	18	24	6	12	30
Round 1	1	1	1			
Residuals	11	17	23	6	12	30
Round 2		1	1	1		
Residuals	11	16	22	5	12	30
Round 3		1	1		1	
Residuals	11	15	21	5	11	30
Round 4	1			1	1	
Residuals	10	15	21	4	10	30
Rounds 5,6,7,8,9,10,11		7	7			7
Residuals	10	8	14	4	10	23
Rounds 12,13,14,15,16	5		5			5
Residuals	5	8	9	4	10	18
Rounds 17,18,19,20,21,22			6		6	6
Residuals	5	8	3	4	4	12
Rounds 23,24,25	3	3				3
Residuals	2	5	3	4	4	9
Rounds 26,27		2		2		2
Residuals	2	3	3	2	4	7
Rounds 28,29		2			2	2
Residuals	2	1	3	2	2	5
Rounds 30,31	2		2			2

Residuals	0	1	1	2	2	3
Round 32				1	1	1
Residuals	0	1	1	1	1	2
Round 33		1			1	1
Residuals	0	0	1	1	0	1
Round 34			1	1		1
Residuals	0	0	0	0	0	0

The PPS(6,3) sampling design produced by this modified procedure is

sample	probability	sample	probability
1 2 3	1/34	3 5 6	6/34
2 3 4	1/34	1 2 6	3/34
2 3 5	1/34	2 4 6	2/34
1 4 5	1/34	2 5 6	3/34
2 3 6	7/34	4 5 6	1/34
1 3 6	7/34	3 4 6	1/34

which is the sampling design in Example 2.2.

References/Bibliography

Asok, C. and B.V. Sukhatme(1976). On Sampford's procedure of unequal probability sampling without replacement. J. Amer. Statist. Assoc. 71 912-918.

Bayless, D.L. and Rao, J.N.K.(1970). An emperical study of stabilities of estimators and variance estimators in unequal probability sampling (n = 3 or 4). J. Amer. Statist. Assoc. 65 1645-1668.

Brewer, K.R.W.(1963). A model of systematic sampling with unequal probabilities. Aust. J. Statist. 5 5-13.

Brewer, K.R.W. and Undy, G.C.(1962). Samples of two units drawn with unequal probabilities without replacement. Aust. J. Statist. 4 89-100.

Durbin, J.(1953). Some results in sampling theory when the units are selected with unequal probabilities. J. Roy. Statist. B 15 262-269.

Durbin, J.(1967). Design of multi-stage surveys for the estimation of sampling errors. Appl. Statist. 16 152-164.

Fellegi, I.(1963). Sampling with varying probabilities without replacement:rotating and nonrotating samples. J. Amer. Statist. Assoc. 58 183-201.

Grundy, P.M.(1954). A method of sampling with probability exactly proportional to size. J. Roy. Statist. Soc. B 16 236-238.

Hájek, J.(1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann. Math. Statist. 35 1491-1523.

Hansen, M.H. and Hurwitz, W.N.(1943). On the theory of sampling from finite populations. Ann. Math. Statist. 14 333-362.

Hanurav, T.V.(1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum, J. Roy. Statist. Soc. B 29 374-391.

Hartley, H.O. and Rao, J.N.K.(1962). Sampling with unequal probabilities without replacement. Ann. Math. Statist. 33 350-374.

Horvitz, D.G. and Thompson, D.J.(1952). A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47 663-685.

Midzuno, H.(1952). On the sampling system with probability proportional to sums of sizes. Ann. Inst. Statist. Math. 3 99-107.

Narian, R.D.(1951). On sampling without replacement with varying probabilities. J. Ind. Soc. Agric. Statist. 3 169-175.

Raj, D.(1956). Some estimators in sampling with varying probabilities without replacement. J. Amer. Statist. Assoc. 51 269-284.

Raj, D.(1964). The use of systematic sampling with probability proportionate to size in a large scale survey. J. Amer. Statist. Assoc. 59 251-255.

Raj, D.(1965). Variance estimation in randomized systematic sampling with probability proportional to size. J. Amer. Statist. Assoc. 60 278-284.

Rao, J.N.K.(1963). On three procedures of unequal probability sampling without replacement. J. Amer. Statist. Assoc. 58 202-215.

Rao, J.N.K.(1965). On two simple schemes of unequal probability sampling without replacement. J. Ind. Statist. Assoc. 3 173-180.

Rao, J.N.K.(1966). On the relative efficiency of some estimators in PPS sampling for multiple characteristics. Sankhya 28 61-70.

Rao, J.N.K. and Bayless, D.L.(1969). An emperical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. J. Amer. Statist. Assoc. 64 540-549.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G.(1962). On a simple procedure of unequal probability sampling without replacement. J. Roy. Statist. Soc. B 25 482-491.

Sampford, M.R.(1967). On sampling without replacement with unequal probabilities of selection. Biometrika 54 499-513.

Sen, A.R.(1953). On the estimate of the variance in sampling with varying probabilities. J. Ind. Soc. Agric. Statist. 5 119-127.

Stevens, W.L.(1958). Sampling without replacement with probability proportional to size. J. Roy. Statist. Soc. B 20, 393-397.

Yates, F. and Grundy, P.M.(1953). Selection without replacement from within strata with probability proportional to size. J. Roy. Statist. Soc. B 15 253-261.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A070134	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) "Controlled Probability Proportional to Size Sampling Designs"		5. TYPE OF REPORT & PERIOD COVERED Interim
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) A. Hedayat and B. Y. Lin		8. CONTRACT OR GRANT NUMBER(s) AFOSR 76-3050C
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Mathematics, University of Illinois at Chicago, Box 4348, Chicago, Illinois 60680		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F-9769-05
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research INM Bolling AFB Washington D. C. 20332		12. REPORT DATE
		13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report)
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Probability proportional to size sampling design. Controlled sampling.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Any sampling design, d, of size n without replacement based on a finite population of N units or N stata can be formally presented by a pair (S_d, P_d), where S_d called the support of d is any set of subsets of size n each based on the elements of U such that the (set theoretic) union of these subsets, called samples, is U and P_d is a strictly positive probability distribution on S_d. A sampling design is said to be a probability proportional to size, denoted by PPS(N,n) , if the probability that the unit i is being selected in a random

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

sample is proportional to a known positive quantity associated with the unit $i = 1,2,\ldots,N$. The literature of survey sampling offers a PPS(N,n) with S_d consists of all $\binom{N}{n}$ possible samples. Here we give an easily applicable technique for the construction of PPS(N,n) with various support sizes and various probabilities on each support. Such sampling designs are needed for controlled samplings when some samples are undesirable to be chosen or we need to minimize (or maximize) the probabilities of the selection of certain samples.